2002 Command and Control
Research and Technology Symposium

Topic: Modeling & Simulation

Proposed Panel Presentation:

Complementary Methods of Modeling Team Performance

*Jared T. Freeman, Ph.D.
Aptima, Inc.
Washington, DC

James A. Pharmer
Naval Air Warfare
Center Training Systems
Division
Orlando, FL

Christy Lorenzen
Micro Analysis & Design, Inc.
Boulder, CO

Thomas P. Santoro, Ph.D.
Naval Submarine Medical
Research Lab
Groton, CT

David Kieras, Ph.D.
EECS Dept., Univ. of Michigan
Ann Arbor, MI

*Contact: Jared Freeman, Ph.D.
Aptima, Inc.
1030 15th St., NW, Suite 400
Washington, DC 20005
202-842-1548 x316 (voice)
202-842-2630 (fax)
freeman@aptima.com

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2002		**00-00-2002 to 00-00-2002**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Complementary Methods of Modeling Team Performance	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Aptima Inc,600 West Cummings Park Suite 3050,Woburn,MA,01801**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**14**	

Complementary Methods of Modeling Team Performance

Jared T. Freeman, Ph.D.
Aptima, Inc.
1030 15th Street
Washington, DC 20005
freeman@aptima.com

James A. Pharmer
Naval Air Warfare Center
Training Systems Division
12350 Research Parkway
Orlando, FL 32826
PharmerJA@navair.navy.mil

Christy Lorenzen
Micro Analysis & Design,
4949 Pearl Circle East
Suite #300
Boulder, CO 80301
clorenzen@maad.com

Thomas P. Santoro, Ph.D.
Naval Submarine Medical
Research Lab
SUBASE NLON BOX 900
Groton, CT 06349
Santoro@nsmrl.navy.mil

David Kieras, Ph.D.
EECS Dept., Univ. of Michigan
Ann Arbor, MI
kieras@eecs.umich.edu

Abstract

Computational tools and techniques for modeling team performance have advanced significantly in recent years. However, there have been few efforts to combine complementary modeling approaches. In the Manning Affordability Initiative, we have applied three modeling technologies to experimental data from a single domain (air defense warfare), a single scenario, and common watchstation technologies (current AEGIS technology and an advanced prototype). The conclusion of this multi-year project in early 2002 offers an opportunity to review the findings. The proposed panelists will summarize a human-in-the-loop experiment conducted to provide modeling data and present findings from efforts to integrate three modeling approaches for design and design validation. Team Optimal Design (TOD) focuses on team modeling. The Integrated Performance Modeling Environment (IPME) uses a general task modeling technique that applies well to individuals or teams. The GOMS Language Evaluation and Analysis Tool (GLEAN) combines individual models of users interacting as a team.

Introduction

The completion this year of the Manning Affordability Initiative (MAI), sponsored by the Office of Naval Research, offers the military R&D community an opportunity to assess a human-centered approach to designing and manning Navy systems. The project consisted of three initiatives. Prototype watchstations were developed using human-centered design methods to enable watchstanders to access and control resources with far greater efficiency than in the past. Computational techniques and tools were applied to model human performance using these systems at the level of the team, the task, and the keystroke. Finally, human-in-the-loop experiments were conducted in which experienced Navy operators executed demanding tactical scenarios using current and advanced watchstations. This paper describes the human performance models and human-in-the-loop testing employed in MAI.

Human-in-the-Loop Experimentation

The goals of the AEGIS comparison study were two-fold. The first goal was to demonstrate that human-centered design of an advanced watchstation could support manning reduction while maintaining or improving performance and maintaining manageable workload levels for individual watchstanders. Second, the comparison provided human-in-the-loop data to support and validate human performance modeling efforts for both watchstation and reduced team design.

Two groups of air defense warfare teams were compared in this investigation. First, in order to evaluate the impact of reduced manning and advanced watchstation design, it was crucial to set the bar high by comparing to intact teams using current technology aboard ship. Consequently, as a benchmark of current performance, data were collected from 8 intact air defense warfare teams composed of eight individuals each. These teams were tested aboard ship pier-side using their own equipment performing an intermediate to advanced level scenario, which was a modified version of a currently used training scenario. The performance of these teams was then compared to reduced-sized teams performing the same scenario but supported by the advanced watchstations. These six intact teams, composed of four individuals each, were tested in the Integrated Command Environment (ICE) Laboratory at the Naval Surface Warfare Center Dahlgren Division (NSWCDD).

While it may be tempting to compare performance on a few critical aspects of a scenario, in a complex domain like air defense a multi-dimensional approach is required to fully understand the effects of watchstation design and manning reduction. As such, teams were compared across three general areas: performance outcomes, workload, and situation awareness and assessment

To compare performance outcomes, timeliness and accuracy data were collected on team performance across a range of detect-to-engage actions for 25 contacts of interest within the scenario. Each of these contacts of interest was carefully embedded into the scenario to investigate performance across a variety of events, which might be encountered within the domain.

Workload was manipulated within-groups by dividing the scenario into two equal segments. The first half of the scenario was considered low difficulty in that there were fewer contacts, which were relatively low threat, and easily identifiable. In the second, higher difficulty segment of the scenario, teams dealt with more tracks, which were more threatening and more ambiguous. Expert evaluator assessments of workload, rated on a scale of 1 to 7, were collected for each watchstander at ten-minute intervals during the approximately 2-hour scenario. In addition, watchstanders provided subjective estimates of their own workload. This was accomplished by administering a modified version of the NASA Task Load Index (NASA TLX) at the end of each scenario half. This index asked each team member to rate their own workload during the previous period on a 20-point scale across ten workload dimensions.

Situation awareness and assessment was measured via online probes, offline questionnaires, and through performance-based inference. The online probes were embedded into the scenario communications that asked specific team members for their contacts of interest, their assessment of the intent of specific contacts of interest and their intent with respect to the contact. Offline questionnaires were administered after each period and queried watchstanders on their contacts of interest during the previous period and why they felt the contact was important. Finally, analysis of performance on several

tactical and strategic actions for specific scenario events provided indicators of situation awareness and assessment.

A detailed description of the results of this investigation is beyond the scope of the current paper. However, it can be said that comparisons across each of these measures showed either comparable or improved performance for reduced-size teams supported by advanced watchstation technology. Overall, the data clearly demonstrated that the incorporation of human factors into the design process has the potential to support manning reduction goals.

In addition to providing objective evidence of the importance of applying human-centered design to support optimal manning goals, the data were also used to support the development and validation of human performance models, which can be used to further support the design process. The remainder of this paper is devoted to discussing how each of the modeling efforts that were executed under the Manning Affordability program worked together and how the human in the loop data that were collected in this investigation provided the opportunity to calibrate and validate these models.

Human Performance Modeling

Data from the human-in-the-loop experiment were used to populate three human performance models.

The models played complementary roles in the project. GLEAN was used early on in the MAI to generate estimates of task times for an air defense warfare operator using a notional interface. Time estimates were passed into corresponding task descriptions in the Integrated Performance Modeling Environment, which included other, SME-defined tasks from the detect-to-engage sequence. IPME was used to develop a broader model of the air defense operation. The IPME model was aggregated for use in the Team Optimal Design system (TOD) to generate and evaluate designs for human teams to staff the air warfare system. IPME was then used to validate that team design. All three models were calibrated and validated using data from the human-in-the-loop experiment cited above.

Team Optimal Design (TOD)

Team Optimal Design is a methodology and associated software that helps acquisitions specialists and designers of man-machine systems to quantify complex aspects of team performance and to perform trade-off analyses that systematically vary team size, the capabilities of team members, their responsibilities, their technologies, mission demands, and other factors. More specifically, TOD helps to answer these questions:

- What is the potential of a baseline team (e.g., the current, fielded team) to accomplish its mission, as measured by mission execution speed, efficiency of coordination, and workload distribution?
- How much does optimal assignment of tasks (alone) improve performance?
- How much do optimal task assignment and optimal team size combined improve performance?
- How much does improved backup potential between team members improve performance?
- How much does broadening the skill of team members – using new technologies or training – improve performance?
- How much does increasing depth of skill or workload capacity – using new technologies or training – improve performance?
- How intense a scenario can the baseline or optimal team execute?
- How reliable is the baseline or optimal team?

Answers to these questions can help designers to determine whether a proposed

technology will have a dramatic effect on team performance, or a minor one. They can help acquisitions specialists assess the benefits of a novel team or system – benefits such as increased mission tempo – given known costs of recruiting, educating, equipping, and supporting its members.

TOD takes as input data that characterize the events of a design reference mission, the tasks by which the man-machine system responds to those events, the capabilities of team members, and, optionally, their responsibilities. TOD then applies multi-objective optimization and clustering algorithms to simultaneously satisfy three objectives: rapid mission tempo, balanced instantaneous workload, and balanced aggregated workload (Levchuk, et al., 2000, 1999). The algorithms can be run repeatedly for teams of different sizes to optimize team size.

TOD's quantitative output, in tabular or graphical format, includes a measure of mission tempo, operator task assignments, a mission execution schedule, a coordination (or communication) schedule, task and estimates of workload and task load.

Aptima applied TOD in two major TOD modeling cycles for the Manning Affordability Initiative. Both cycles were executed with the support of a program-wide work group consisting of domain experts, modelers, and experimentalists. The first modeling phase, completed in 1999, implemented a complex model of task flow in AAW operations using an advanced prototype, a prototype command and control technology. The model also represented constraints on task assignment, some of the most influential of which concerned coverage of communications circuits and track authority. The level of abstraction at which the model was cast and the fidelity of its parameter values enabled the researchers to design a new team for the prototype that had roughly the same task assignments as one generated independently by domain experts.

The second TOD modeling cycle, completed in 2001, was designed to leverage empirical data concerning human execution of a small set of tasks – from track detection through engagement – in an AAW scenario. The effort consisted of three tasks, reported below.

The first task produced a TOD model representing AAW operations in a team using AEGIS equipment. It was calibrated against experimental data (generated by AEGIS operators) to ensure that it produced accurate estimates of workload and task latency. This was a verification of TOD, a test of its internal validity in which empirical data were input, the output was compared to empirical observations, and the model was iteratively refined.

The second task produced a TOD model representing a smaller AAW team executing the scenario using the prototype. It, too, produced estimates of workload and task latency, which NAWCTSD compared to empirical data to independently assess the validity of the model.

In the third task, we embarked on a series of modeling excursions to demonstrate the rapidity and utility of using TOD to contrast alternative systems and team architectures.

AEGIS Model

The task flow model of AEGIS AAW operations consisted of the fourteen tasks – derived from the Detect-to-Engage (DTE) sequence – on which empirical data were collected by NAWC/TSD. The model (represented by Figure 1~~Figure 1~~Figure 1~~Figure 1~~) had two sections, one representing detection and identification activities, the other representing explicit actions to query, deter, or attack a track.

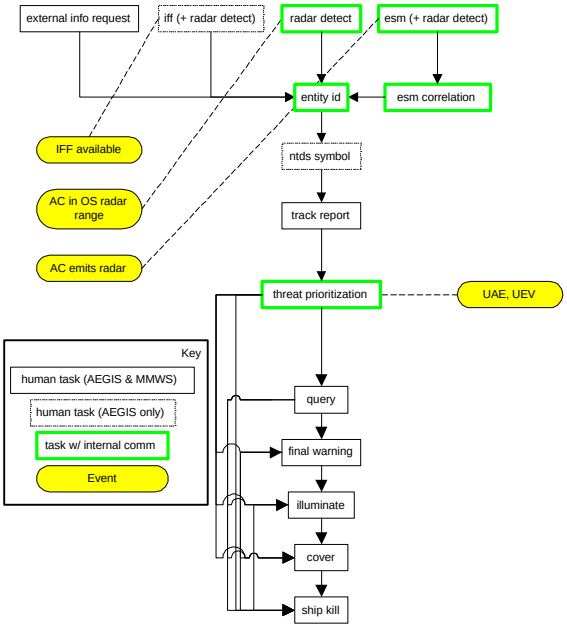

Figure 1: Task flow for AEGIS & Prototype.

Virtually all items in the lower section were interconnected, indicating that the AAW team has the latitude to respond to events using virtually any combination of actions. Omitted from this diagram and from the TOD AEGIS model were tasks that could not be reassigned in TOD optimization between AEGIS and the prototype, including air control tasks (DCA escort).

These tasks in the AEGIS TOD model were parameterized with duration (mean and SD) and workload values. Parameter values were estimates developed by the work group and implemented in IPME. These values were refined in consultation with the work group to account for differences in the meaning of tasks between IPME and the TOD model. IPME decomposes workload into visual, auditory, cognitive, and psychomotor load. TOD represents workload for a given task using a single value. This value was computed as the maximum of IPME workload values. Several tasks in the model involved external communications. For those tasks, we lengthened raw task duration by one second per 3.4 words. The standard deviation of duration for these tasks was lengthened proportionately. Internal communications were represented as delays between sequential tasks that were executed by different individuals. The length of each delay was four seconds, equivalent to the duration of the average, 12-15 word internal comm.

The task flow model allows for many possible responses (sequences of tasks) to any given scenario event. The few that were modeled were derived from the empirical data. Tasks were assigned to operators using an AEGIS organizational architecture of eight people.

The results of AEGIS modeling conformed closely to the empirical values. We are restricted from reporting detailed results here. However, latencies from this model were within the predicted limits, and average workload was very close to participants' subjective estimates.

Advanced Technology Model

The data and runtime procedures used in modeling the advanced technology were identical to those used in AEGIS modeling with the following exceptions.

- Two tasks that were fully automated under the advanced technology were removed from the model.
- Durations and workloads for other technologically supported tasks were reduced per SME estimates.
- Tasks were assigned to four operators per the organizational architecture for this system.

These changes in parameter values produced a significant decrease in latencies due to full and partial automation of tasks as well as decreases in coordination requirements between members of the (reduced) operating staff.

Excursions on the Advanced Technology Model

TOD's value to designers lies in its ability to replicate a baseline condition and then to rapidly generate alternatives to that baseline team. To demonstrate this functionality, we conducted a series of excursions from the AEGIS model. Specifically, we tested the effects on mission tempo of manipulating task assignment, team size, operator capacity, breadth of expertise, and task backup capacity. The excursions were performed in a matter of several hours, indicating the rapidity with which TOD can generate designs once a task flow model, a representative scenario, and accurate task parameters are in place.

The excursions varied several parameters of the baseline AEGIS model:

- The number of operators was varied from eight (in the AEGIS model) to as few as four.
- The maximum workload capacity of seven units per operator in AEGIS was raised as much as 30%. This parameter can be interpreted as depth of expertise or of system support for task execution.
- The AEGIS maximum of eight tasks per operator was raised as high as 12 in the excursions. This parameter represents the breadth of expertise of operators.
- A maximum of five tasks in common between operators in AEGIS was lowered as far as zero. This parameter represents backup potential between operators.

The results, relative to the AEGIS model, were as follows:

- Optimizing task assignment alone significantly accelerated mission tempo.
- Reducing team size & optimizing task assignments accelerated tempo provided that the team had at least five operators.
- Increasing expertise or boosting it technologically improved performance.
- Reducing team size, optimizing assignments, and increasing expertise enabled four operators to perform as well as eight.
- Increased breadth of skill plus optimized task assignment had little impact over optimized assignment alone.
- Reducing backup capacity and optimizing task assignment produced minimal decrements in performance. Specialization of operators was feasible.

In sum, the Team Optimized Design software and associated techniques produced accurate models of empirical data. Rapid excursions on the AEGIS model illustrated TOD's utility for acquisitions professionals and designers who are challenged with quantifying the effects of team size, task assignment options, work backup schemes, training for depth, and cross training.

Integrated Performance Modeling Environment (IPME)

The IPME is a discrete event simulator that uses task sequence as the primary organizing structure. The modeling process involves task analytic decomposition of human behavior, from larger units to successively smaller elements of behavior, until a level of reduction is reached that can provide reasonable estimates of human performance for task elements.

Users of IPME software can model environment variables, operator traits and states, performance modifiers and dynamic crew assignment. Workload assessment and specialized experimental human performance schedulers are built into the IPME. These include Prediction of Operator Performance, the Information Processing / Perceptual Control Theory scheduler, VACP and W/Index. Multiple runtime model integration capability allows linkage to external client or server simulations.

One role of IPME simulation within the Manning Affordability Initiative was to

demonstrate the utility of human performance modeling to the design process. The modeling effort followed the following process to insure the accuracy and validity of the models and demonstrate the capability of the models to provide useful predictions for designers of advanced technology to support optimal manning (Scott-Nash, Carolan, Humenick, Lorenzen, & Pharmer, 2000).

1. Model performance and workload of air defense warfare teams using current watchstation technology on a two-hour intermediate-to-advanced level scenario
2. Using the same scenario, collect human in the loop data from intact teams aboard ship using this current technology
3. Validate the 'baseline' model with the experimentally collected data, and, if necessary, calibrate the baseline model
4. Modify the model to reflect warfighter-centered design changes to the watchstations and manning reductions achieved under MAI comparison study.
5. Again using the same ADW scenario, collect performance data from these smaller teams using the warfighter-centered design watchstations
6. Using the modified model, predict changes in human performance and workload trends with the new watchstation and reduced team size
7. Validate these predictions against the collected data.

Five key measurements representing major ADW activities were selected as the criteria for validating the models. These included:

- Average Time to First ID
- Average Time to New Track Report
- Number of Tracks Queried
- Number of Tracks Warned
- Number of Tracks Identified

In addition to these key performance parameters, the modified IPME model of air defense warfare performance was capable of estimating differences in workload between the 'baseline' performance and performance of reduced teams using advanced watchstation technology. The results of the validation indicate reasonably strong agreement with a number of the performance measures as well as a strong agreement (often within 25%) with the empirical workload measures for the advanced watchstation teams (Scott-Nash, Brockett, & Pharmer, 2001).

An important conclusion of this effort was that it is possible to model such large and complex processes. One advantage of creating such a model is that now a valid ADW model exists that can be used again and again for various purposes. For example, further analysis on possible design or crew changes can be explored at very little additional cost.

GOMS Language Evaluation and Analysis tool (GLEAN)

The GOMS (Goals, Operators, Methods, and Selection Rules) methodology (Card, Moran, and Newell, 1983; John & Kieras, 1996a, b) for describing human procedural knowledge in a programming-language-like notation is among the most common engineering models in use for interface design. While well known for its capacity to deal with fine details of the HCI and associated human behaviors, it has, however, rarely been applied to model and predict the performance and outcomes of team activities rather than activities of individuals working on their own.

GOMS models of watchstanders in an air defense warfare team have been developed using GLEAN, the GOMS Language Evaluation and Analysis tool, created by Dr. David Kieras of the University of Michigan (Kieras et.al., 1995; Kieras, 1998). Communications and collaborations among model team members are facilitated by a dynamic interrupt mechanism. The sensory modality processors in GLEAN can generate interrupts to activity in the cognitive processor

in order to insert volatile information from sensory memory into cognitive working memory.

For example, the auditory processor can be primed to listen for keywords in ongoing voice communications. When a keyword occurs, an interrupt is triggered to load related information into memory store where it can be accessed by cognitive processes. Analogous visual behavior is problematic, however, as vision is used in conjunction with cognitive processes where items on a display or in a table are held in focus causing events outside of the field of view to be missed and therefore not available to trigger an interrupt. Thus the spontaneous capture of all critical visual events is considered as the upper limit to expected visual performance and more realistic models of deliberate visual search are used to probe for the lower limit in order to bracket expected performance. However, the automatic interception of auditory events via interrupts is a reasonable model of auditory behavior and essential to spontaneous inter-operator voice communications. Since spontaneous communication is critical to team collaboration and workload sharing, this interrupt mechanism is the key to using a team of GOMS models for studying these behaviors in a team of humans.

Alternative Strategies for Visual Search

The representation of human visual search for new and changed track icons on a tacsit display is a difficult modeling and simulation problem that can and does tax the capabilities of much more complex modeling approaches than GOMS. The GOMS philosophy toward such problems is to bracket expected performance by posing hypothetical best case and worst case task execution strategies (see Kieras & Meyer, 2000). This allows the estimation of high and low limits for very complex behaviors like visual search with simple combinations of elementary behaviors that have been well defined. When visual events are initially inserted into the GLEAN synthetic environment the information that they are "new" has a lifetime of 0.2-0.5 seconds. The upper limit for search performance obtains when all visual events are detected during that brief lifetime and logged into working memory. This condition, termed universal interrupts, affords the model the maximum possible time to evaluate the threat and decide on actions as the exercise evolves. Any search strategy that does less than this universal coverage of critical visual events would be subject to increasing likelihood of missing events or responding to them too late to perform important actions.

More realistic visual search techniques would involve deliberate start and stop of the search process at selected points during or after execution of other tasks where vision is occupied. This would assume that, when the operator is engaged in locating or reading visual information for a particular task from another screen window, the visual system is not capable of simultaneously detecting events from a tacsit display. A conservative lower limit for the worst visual search performance would be the case where only the final range tripwire event would be captured, i.e. the model only notices hostile tracks when they come into such close proximity with ownship that they pose immediate danger. These two assumptions then bracket the range of expected visual search performance.

Sensory-Motor and Cognitive Workload Measures

In modeling human performance with the IPME task network tool, it is necessary to describe discrete tasks in terms of their relative distribution of work in sensory-motor and cognitive modalities. In general such task workload descriptions are estimated by individuals familiar with the subject matter area who have performed related tasks.

Since the GLEAN tool simulates the activity of sensory-motor processes using time

parameters derived from experimental psychology, it offers the possibility of estimating not only the complete time duration of a given task, but also the portions of that duration during which the different modalities are active. For example, statistics are recorded on the total number and duration of visual actions performed on each repetition of each task. The overall totals for any designated time period can also be computed as desired for any individual operator model. By combining estimates from the different modalities and scaling the numbers in various ways, workload predictions can be made and their correspondence to subjective workload estimates can be determined.

Predicting Workload

In order to correspond with the observer estimates from the human-in-the-loop data, the totals for workload measures were computed at 10-minute intervals for each GOMS model over the test scenario and scaled into the range of 1 to 7 used by the observers. Workload values were scaled against the maximum recorded value over the exercise.

The workload observers were selected for their expertise in the tasks of a particular watchstander and were instructed to base their estimates on the activity level of that operator relative to their estimate of the individual's maximum possible work output. At this time, GLEAN does not provide an equivalent maximum workload capacity of a simulated operator, and so the scaling to the 1-7 range was done simply using the maximum and minimum values produced by the model. In mitigation, it is not clear how reliable or valid the observer's subjective estimate of maximum capacity would be. In addition, because observations were limited to a single individual per observer, it is difficult to determine the level of agreement and reliability of subjective workload estimates across the observers.

Verification of Model Team Configurations

GOMS models of operators in an anti-air warfare exercise using watchstations with advanced human-computer-interface technology were constructed and a set of GOMS Methods was written to perform the major air warfare tasks in the scenario. Using data collected from team performance on the low-difficulty first half of the scenario, an acceptably accurate model was found by systematically starting from two bracketing models and developing a model that matched the data acceptably well. This calibrated model was then validated against performance data collected during the higher difficulty second half of the scenario. The calibration process, known as the "Verification" portion of the "Verification and Validation," or V&V process, involved a number of iterative model cases as follows. In the first case, the Methods were assigned to three separate operator models corresponding to three members of the five-member human team. The three operators chosen for modeling were the ones with primary responsibility for the required actions involving threat air tracks in the scenario.

Three-station-universal-search. In the first case, each operator worked independently on separate tasks without verbal communications or any collaboration with the other two team members. In addition, visual search was assumed to occur in parallel with all other activities and capture all critical visual events. Under this condition, no track appearance or change event was missed by any of the three model operators, all appropriate actions on critical tracks were taken, and the time latencies of the actions were similar to, or, in a number of cases, shorter than, the fastest times produced by the human operators. This model represents the best performance, corresponding to the upper bracket, but it involves an unrealistic ability to reliably detect all of the significant visual changes in the display.

Three-station-deliberate-search. The next case attempted to bracket the worst expected performance limit by using the same no-communications model with the additional restriction of brief 3-5 second deliberate search time windows occurring between threat-related activities. Only track appearance and change events that happened to fall within these windows would be captured by this model. Again, the model operators did not collaborate through verbal or any other communication mode. This model resulted in the poorest fit of predicted to actual latencies for actions taken and in addition had the highest number of missed required actions of all models tested. The average error statistic for the workload fit was also relatively high. This model was our worst-case, or lower bracket, model.

Models with Voice Communications

Further models were built to bridge between the upper and lower bracket cases through the introduction of various hypothetical communications between the modeled team members. The actual human teams freely communicate over their internal network depending on different individual styles. Many ad-hoc remarks are made about track events and various pieces of information are passed. It is not at all clear from a study of these communications to what extent they are useful to, or used by, their respective recipients. Hence our approach was to propose that certain information was communicated within the team and then determine whether adding that capability to the model would improve performance towards the best case. We built two variations of a three-member model team with voice communications and two variations of a four-member team with voice communications. The model iterations were stopped at a four-member team in which two members provided threat track search and identification assistance to the other two members. This model succeeded in accomplishing all the actions on critical tracks that were made by the actual teams whereas each of the other models missed a few required actions**.**

Model Validation

Once a team model was built that performed acceptably on the Verification data, which was from the first half of the test scenario, the model was used to predict the second half data set from the same exercises for Validation. The overall latency and workload predictions for this model were reasonably close to the data, in several cases being within ten percent, a common rule of thumb for engineering design purposes. The results for prediction of overall workload observations in ten-minute intervals on the Validation section of the exercise are shown in Figure 2~~Figure 3~~ ~~Figure 3~~ ~~Figure 2~~.

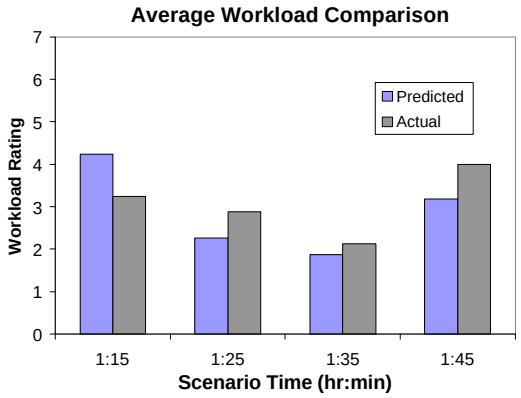

Figure 2~~3~~~~3~~2: GOMS workload validation. The team average predictions do match the observed averages for each interval quite well.

Conclusions

The good fit of the workload predictions is encouraging for the prospects for further development of workload prediction with a GOMS tool. The fact that team voice communications were valuable mainly to help with the visual detection problem suggests that team performance could be improved directly by improvements in the workstation design, which would then leave team communication channels free for other, more complex, team

activities. Also, the key role of the voice communications suggests that further development of accurate models for this processing would be valuable. Finally, this work demonstrates that the concept of modeling a team of humans with a team of models of individual humans is a viable approach to bridging the gap between the psychology of individual humans and the organization and functioning of teams.

Acknowledgments

This work is funded by the Office of Naval Research, supported by the Naval Sea Systems Command, and managed by the Naval Air Warfare Center Training Systems Division. We are obliged to the men and women of the Navy who participated in experimental studies, and to the subject matter experts who assisted the modeling team. The opinions expressed here are the authors' and do not necessarily reflect the views of the U.S. Navy or the Department of Defense.

References

Card, S., Moran, T.P., and Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.

John, B. E., & Kieras, D. E. (1996a). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction*, 3, 287-319.

John, B. E., & Kieras, D. E. (1996b). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction*, 3, 320-351.

Kieras, D.E. (1998). *A guide to GOMS model usability evaluation using GOMSL and GLEAN3*. (Technical Report No. 38, TR-98/ARPA-2). Ann Arbor, University of Michigan, Electrical Engineering and Computer Science Department. January 2, 1998. Current version available online at ftp://www.eecs.umich.edu/people /kieras /GOMS/GOMSL_Guide.pdf.

Kieras, D.E., Wood, S.D., Abotel, K., & Hornof, A. (1995). GLEAN: A Computer-Based Tool for Rapid GOMS Model Usability Evaluation of User Interface Designs. In *Proceeding of UIST*, 1995, Pittsburg, PA, USA. November 14-17, 1995. New York: ACM. pp. 91-100.

Kieras, D. E., & Meyer, D. E. (2000). The role of cognitive task analysis in the application of predictive models of human performance. In J. M. C. Schraagen, S. E. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum, 2000.

Levchuk, Y. N., Luo J., Levchuk G. M., Pattipati K. R., and Kleinman, D. L. (1999) A multi-functional software environment for modeling complex missions and devising adaptive organizations. *Proceedings of the 1999 Command & Control Research & Technology Symposium*, War College, Newport, RI, June/July 1999.

Levchuk, Y.N., Pattipati, K.R. and Kleinman, D.L. (1999). Analytic model driven organizational design and experimentation in adaptive command and control. *Systems Engineering*, Vol. 2, No. 2, 1999.

MacMillan, J., Paley, M.J., Levchuk, Y.N., Entin, E.E., Serfaty, D., and Freeman, J.T. (In press). Designing the Best Team for the Task: Optimal Organizational Structures for Military Missions. In Mike McNeese, Ed Salas, and Mica Endsley (editors), *New Trends in Cooperative Activities: System Dynamics in Complex Settings*. San Diego, CA: Human Factors and Ergonomics Society Press.

Scott-Nash, S., Carolan, T., Humenick, C., Lorenzen, C., Pharmer, J. (2000)Calibrating and validating a human performance model to support predictions

of future military system capability, *Proceedings of the Interservice / Industry Training, Simulation and Education Conference*, Orlando, FL

Scott-Nash, S., Brockett, C. H., & Pharmer, J. A.(2001). Verifying and validating the AEGIS air defense warfare model. *Proceedings of the Interservice Industry Training Simulation and Education Conference (I/ITSEC)*. Orlando, Florida.